FIG. 1

SQL
REQUEST

130

SESSION CONTROL

— 200

PARSER

— 205

DISPATCHER

— 210

FIG. 2

SQL
REQUEST

205

INTERPRETER

300 —

SYNTAX CHECKER

305 —

SEMANTIC CHECKER

310 —

DATA DICTIONARY CHECKER

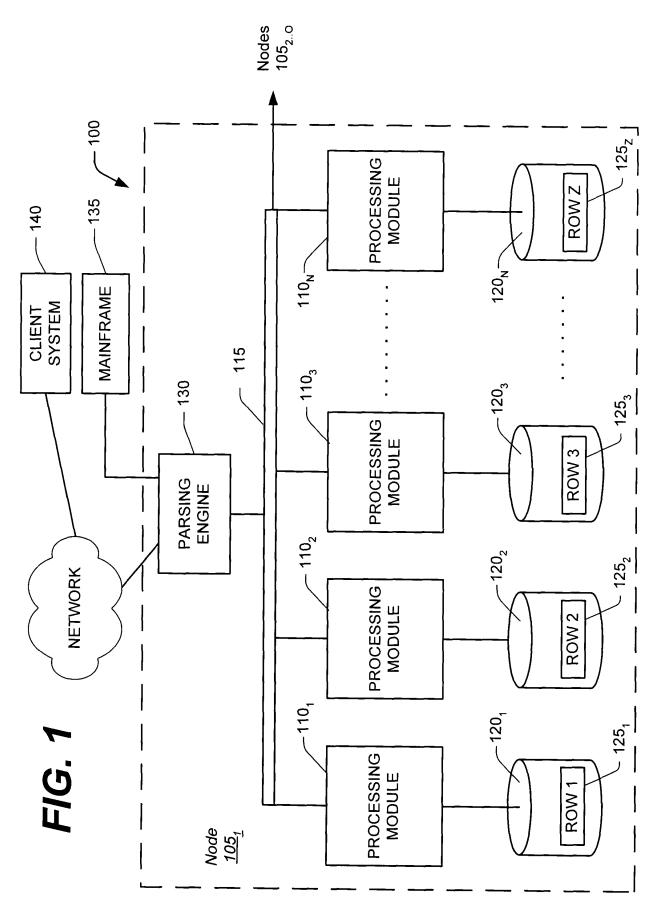315 —

OPTIMIZER

320 —

EXECUTABLE
STEPS

FIG. 3

Express Mail #EV339222559US
Filed: 09/12/03; File No. NCR 11092
Inventor: Ramesh, et al.
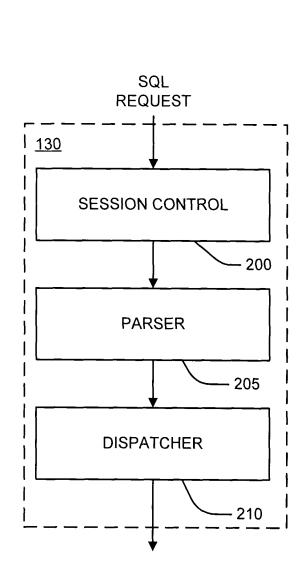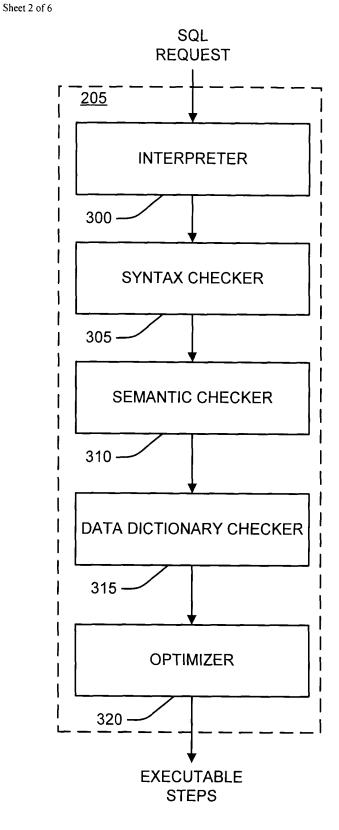For: Clustering Strings
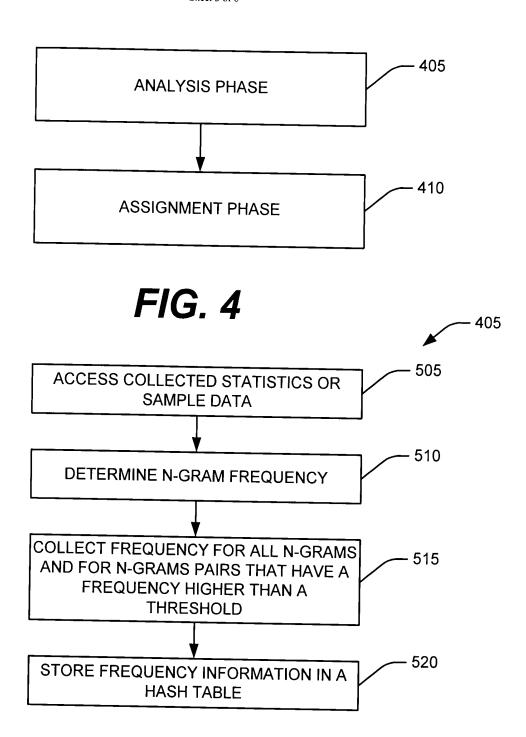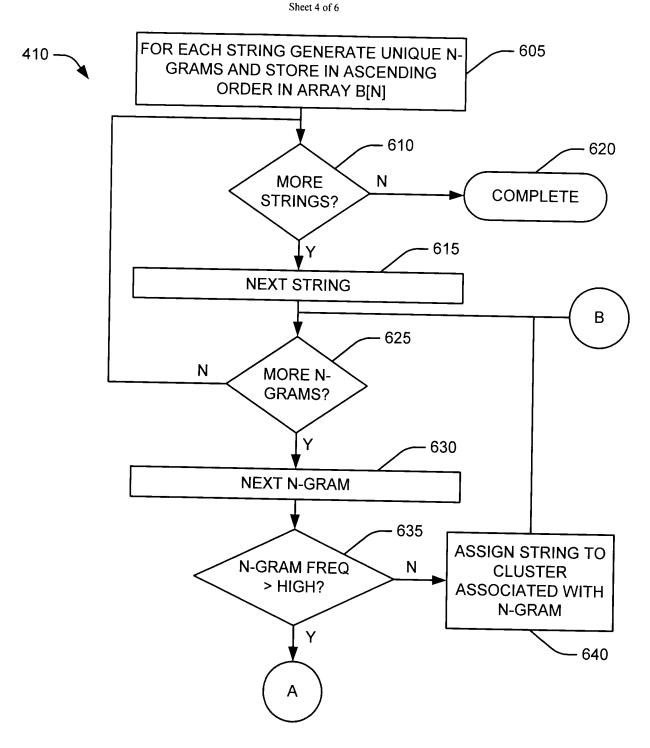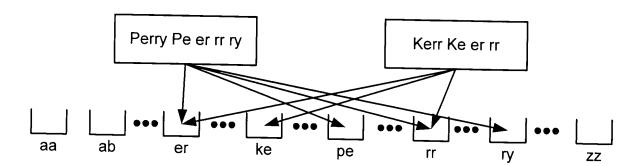Contact: John D. Cowart/858-485-4903
Sheet 3 of 6

```
┌─────────────────────────────┐
│                             │── 405
│       ANALYSIS PHASE        │
│                             │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│                             │── 410
│      ASSIGNMENT PHASE       │
│                             │
└─────────────────────────────┘
```

FIG. 4

```
                                        ── 405
┌─────────────────────────────┐
│  ACCESS COLLECTED STATISTICS OR │── 505
│         SAMPLE DATA             │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│  DETERMINE N-GRAM FREQUENCY     │── 510
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ COLLECT FREQUENCY FOR ALL N-GRAMS │
│ AND FOR N-GRAMS PAIRS THAT HAVE A │── 515
│   FREQUENCY HIGHER THAN A         │
│         THRESHOLD                 │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│  STORE FREQUENCY INFORMATION IN A │── 520
│          HASH TABLE               │
└─────────────────────────────┘
```

FIG. 5

410

```
┌─────────────────────────────────────┐
│ FOR EACH STRING GENERATE UNIQUE N-   │──── 605
│ GRAMS AND STORE IN ASCENDING         │
│ ORDER IN ARRAY B[N]                  │
└─────────────────────────────────────┘
```

610 — MORE STRINGS?

N → COMPLETE — 620

Y

615 — NEXT STRING

B

625 — MORE N-GRAMS?

N

Y

630 — NEXT N-GRAM

635 — N-GRAM FREQ > HIGH?

N → ASSIGN STRING TO CLUSTER ASSOCIATED WITH N-GRAM — 640

Y → A

FIG. 6A

Express Mail #EV339222559US
Filed: 09/12/03; File No. NCR 11092
Inventor: Ramesh, et al.
For: Clustering Strings
Contact: John D. Cowart/858-485-4903
Sheet 5 of 6

A

FOR EVERY OTHER N-GRAM IN THE STRING — 645

N-GRAM FREQ > HIGH? — 650

N-GRAM PAIR IN FREQ MAP? — 655

ASSIGN STRING TO CLUSTER ASSOCIATED WITH N-GRAM PAIR — 670

FOR EVERY OTHER N-GRAM IN THE STRING — 660

ASSIGN STRING TO CLUSTER ASSOCIATED WITH N-GRAM TRIPLE — 665

B

FIG. 6B

Express Mail #EV339222559US
Filed: 09/12/03; File No. NCR 11092
Inventor: Ramesh, et al.
For: Clustering Strings
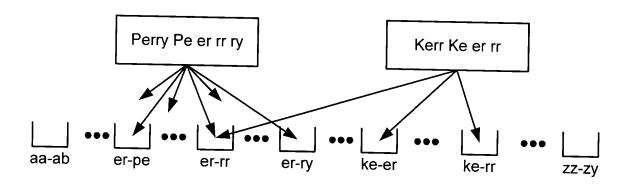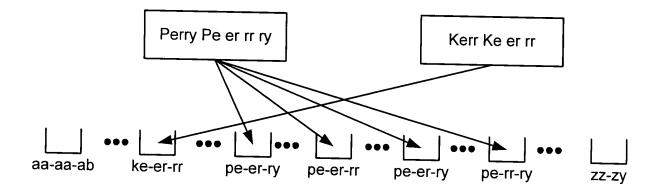Contact: John D. Cowart/858-485-4903
Sheet 6 of 6

FIG. 7

FIG. 8

FIG. 9